Visual Voice Activity Detection Using Frontal Versus Profile Views

Rajitha Navarathna*, David Dean*, Sridha Sridharan*, Clinton Fookes* ands Patrick Lucey†
*Speech, Audio, Image and Video Technology Lab, Queensland University of Technology, Australia.
†Disney Research Pittsburgh, USA.

***Abstract*—Visual activity detection of lip movements can be used to overcome the poor performance of voice activity detection based solely in the audio domain, particularly in noisy acoustic conditions. However, most of the research conducted in visual voice activity detection (VVAD) has neglected addressing variabilities in the visual domain such as viewpoint variation. In this paper we investigate the effectiveness of the visual information from the speaker's frontal and profile views (i.e left and right side views) for the task of VVAD. As far as we are aware, our work constitutes the first real attempt to study this problem. We describe our visual front end approach and the Gaussian mixture model (GMM) based VVAD framework, and report the experimental results using the freely available CUAVE database. The experimental results show that VVAD is indeed possible from profile views and we give a quantitative comparison of VVAD based on frontal and profile views The results presented are useful in the development of multi-modal Human Machine Interaction (HMI) using a single camera, where the speaker's face may not always be frontal.**

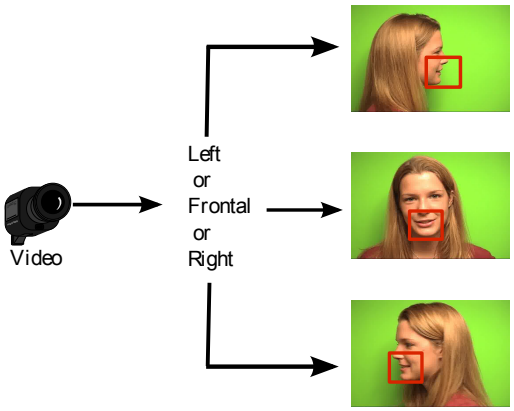


Fig. 1. An example of the setup to capture the frontal, left and right poses. The detected region-of-interest (ROI) was directed to the VAD system, to recognize the speech and non-speech.

I. Introduction

The detection of voice activity (i.e. when speech occurs and not what is said) is a challenging problem, especially when the level of acoustic noise is high. Most current approaches only utilise the audio signal, making them susceptible to acoustic noise [1, 2]. Frame-energy [3] and entropy [4] are some of the audio based techniques which can be used for voice activity detection (VAD). However, the robustness and effectiveness depends on the acoustic environment and these approaches perform poorly when the level of background noise increases. An obvious approach to overcome this problem is to use the visual modality in the form of speaker's lip information as it is not susceptible to the problems associated with audio based VAD.

In visual speech recognition or in lip reading, hidden Markov models (HMMs) [5] are used as the recognition tool and it is widely recognised that this is the defacto standard. However, in terms of detecting visual voice activity, there is no current standard technique being used. This is because research in this area has been rather dormant.

There are few attempts to incorporate the visual modality in VAD. An early work in visual voice activity detection was the work done by Liu and Wang [6], where they presented a visual VAD (VVAD) framework using a template matching method and applied principal component analysis (PCA) [7] for the feature extraction on the detected mouth region. They modeled the distribution of speech and non-speech using two different Gaussian mixture models (GMMs). The authors, demonstrated their experiments using two French speakers. Sodoyer et. al [8] presented a VVAD system using lip contour geometric parameters where they applied temporal smoothing to extract the visual features using the width and the height of the mouth region. In 2007, Libal et. al [9] developed a real-time system to recognise visual speech activity on low cost embedded platforms. This system uses a camera mounted on the rearview mirror to monitor the driver. It detects face boundaries and facial features, and finally employs lip motion clues to recognise VAD. More recently, Aubrey et. al [10] proposed a method for VVAD based on the optical flow of the speaker's mouth region. The authors show that they can obtain less false detection when they train on a small number of observations. A small database was utilised which contains only one male and one female speaker. Furthermore, all the above research work in VVAD has been conducted using only the frontal images with small amounts of data. A realistic scenario of using the visual modality to detect voice activity is to have the system being able to function in different views. An example of this is given in Figure 1. Having a VVAD system which can recognize the voice activity from both frontal and profile views will be a major benefit to many applications such as voice based Human Computer Interaction. For example in-vehicle environments. Research todate on VVAD has been conducted only on the frontal view of the face and no research has addressed VVAD using profile views. In our work we present a VVAD system using different

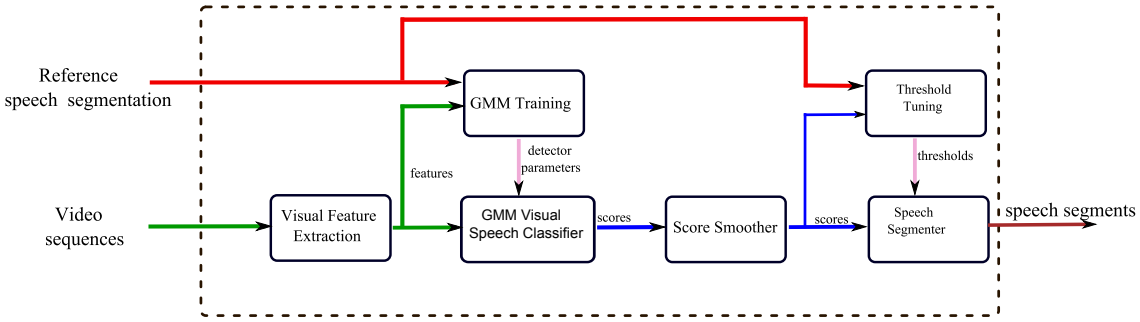


Fig. 2. An overview of the visual voice activity detection framework.

poses using the publicly available CUAVE database [11], which consists of 36 speakers. We view this work as the first and necessary step to develop an efficient human computer interaction system based on audio-visual input.

II. GMM BASED VAD

This section gives an overview of the speech detection framework and the detailed module descriptions in the sub sections. An overview of the operation of the main components of the speech detection framework is outline in Figure 2.

A. Feature extractors

The feature extraction module is designed to extract the visual features from the ROIs. The module is designed to extract the visual features by first dividing the incoming speech sequence utterance into a number of fixed-length frames, at a particular frame-rate, and then return a feature vector for each frame and direct it to the GMM visual speech classifier. A brief description of the feature extraction stage can be found in Section III.

B. GMM visual speech classifier

The Visual speech classifier module was implemented to estimate the speech-likelihood by assigning scores to frames of feature files. Feature files are generated from the feature extractor module. The GMM classifier module takes the feature sequences and produces the corresponding list of scores files. Output score files are a list of scores for each frame of the video utterance. The GMM classifier is trained using the speech/non-speech reference speech segmentation (ground truth values) obtained on the training set. GMM training takes the features corresponding to speech and non-speech events from the video utterance in training set in order to estimate the means and the variation of each Gaussian mixture. Two 8 mixture GMMs were used to separately model speech and non-speech events and classified scores are given as the log likelihood ratio of the speech GMM over the non-speech GMM.

C. Score smoother

The score smoother takes a list of score files from a speech detector module and produces a corresponding list of smoothed score files. For this research, a one-second median filter is used for smoothing. This smoother operates by replacing each score with the median of a one-second window centered on the score.

D. Speech segmenter

The speech segmenter module is the final stage of the framework. This stage converts the log-likilihood score files into speech/non-speech segment decisions. It is designed to take a list of score files and a threshold value and produce a corresponding list of speech segment files which can be compared with the reference speech segment files (ground-truth values) to evaluate the performance of VVAD system. A simple threshold-based segmentation is used where the output of the smoother is divided based upon a single threshold; frames below the threshold are designated *non-speech* and frames above are designated *speech*. The training data was used for the tuning of the segmentation thresholds based on minimising the half total error rate (HTER) defined in Section IV-C.

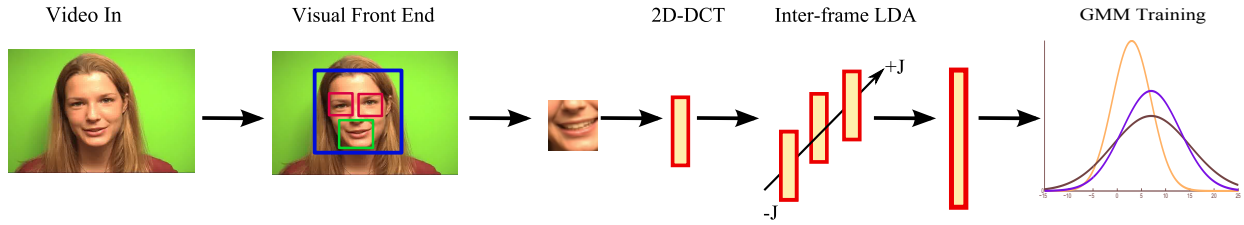
III. VISUAL FEATURE EXTRACTION SYSTEM

A. Visual front-end

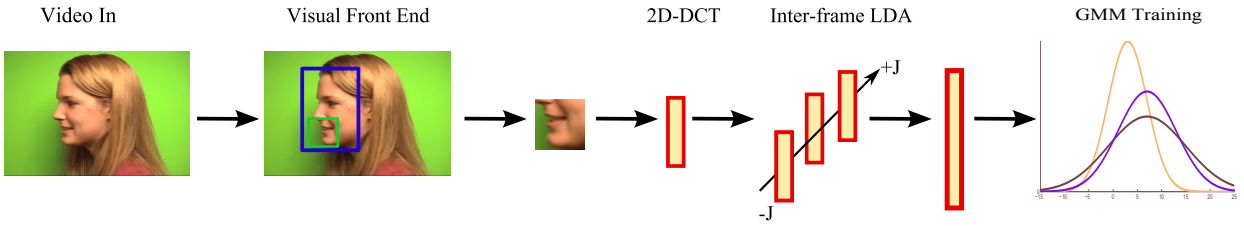
An efficient visual front end system which is able to track and locate the ROI from the speaker's frontal or profile (i.e left and right profile) face and lip area was developed using the Viola-Jones algorithm [12]. The visual-front end was similar to Lucey et. al [13].

Given a video of a speaker, initially face localization is applied according to the view to estimate the position of the speaker's face using 16×16 frontal or profile face classifiers. If the face image is frontal, the eyes were searched over specific regions of the face. Next, the mouth center classifier was used to refine the search region. The resulting mouth region was then used as the search region to locate the right and left mouth corners. After locating the mouth corners, the extracted mouth ROI was rotated so that these two points were aligned horizontally.

The visual front-end for the profile view was similar to the frontal view. Once the face is detected we used a 20×20 eye classifier and a 15×15 nose classifier to localise the profile eye and the nose. The mouth region was located in the bottom part of the face region. Once the general mouth region is found,



(a) From frontal view.



(b) From profile view.

Fig. 3. An overview of the visual feature extraction system.

the left mouth corner is detected. The extracted mouth ROI was normalised based on the left mouth corner.

Finally, the extracted frontal and profile views mouth ROI was smoothed using a mean filter and downsmapled to 40×40 to keep the dimensionality low. This process was performed on every incoming video frame of the speaker. All the classifiers were developed using the OpenCV libraries. An example of extracted mouth ROIs from frontal, left and right views is presented in Figure 4.

B. Visual features

We extracted the visual features in the form of cascading appearance based features, which consist of both static and dynamic feature extraction [14] stages. In the field of audio-visual speech recognition (AVSR) [15], this method has been established as the state-of-the-art for visual feature extraction.

Following the ROI extraction from the visual front end system, an image mean normalization step was performed to remove any irrelevant information, such as illumination or speaker variances. The mean image was calculated from the given entire utterance and subtracted from every incoming frame in the utterance, before extracting the static feature vector. The subtracted image is called the mean removed image (MRI). Then two-dimensional separable, discrete cosine transform (DCT) is applied to the MRI and the top 100 higher energy components were selected to capture the static information.

Visual speech is represented by the movements of the visual articulators. The best features for representing visual speech are generally considered to focus on the movement of the features, rather than the features within each frame. In order to incorporate dynamic speech information, the static features were concatenated before speech-class based linear discriminant analysis (LDA) was performed based on a known transcription.

We used seven of these neighboring static feature vectors over ±3 consecutive frames were concatenated around the frame under consideration, and projected via an inter-frame LDA step to yield a 40-dimensional "dynamic" visual feature vector, extracted at the video frame rate of 30 Hz. The classes used for LDA matrix calculation were the HMM states, based on forced alignment using audio transcriptions. A depiction of the visual feature extraction system for the frontal and the profile views is given in Figure 3.

IV. Experimental Setup

A. Research data

The experiments were conducted using the freely available audio-visual CUAVE database [11], which contains speakers talking in frontal and non-frontal poses. It consists of two sections: the individual and the group section. The individual section was designed to give realistic conditions such as speaker movement, while the group section was included to look at pairs of simultaneous speakers.

The CUAVE database consists of 36 speakers (19 male and 17 female speakers). The database has over 7000 utterances and all the recorded speech was in English. The data were collected using frontal, left and right views. In the frontal view, each speaker spoke 50 digits whilst standing still naturally. In the profile views, each speaker utters 10 digits. Some of the examples of the various speakers and poses available in the CUAVE database are given in Figure 5.

B. Evaluation protocol

The main motivation behind the creation of the CUAVE database was to create a flexible, realistic and easily dis-

(a) From left profile. (b) From frontal view. (c) From right profile.

Fig. 4. Examples of the extracted 40×40 ROI Images using different views.

Fig. 5. Examples of the CUAVE individual sequences: (a) Left profile, (b) Frontal view and (c) Right profile.

tributable database that allows a representative and fairly comprehensive testing in speaker-independent audio-visual speech recognition not for VVAD. Therefore we derived an evaluation protocol for task of VVAD using the CUAVE database for our experiments.

An example of the categorisation of the speech and non-speech events for the frontal and profile views is shown in Figure 6. We categorised the entire "zero" to "nine" session as *speech* and the session between "nine" to "zero" as *non-speech* in the frontal view. In the profile view, we categorised the digits as *speech* and silence between digits as *non-speech*. The main reason was there were significant amounts of silence between digits in the profile data in most of the subjects. For example, as shown in Figure 6(b), there are 10 digits around 25s, but there are around 30 digits in Figure 6(a). We selected 24 subjects from the CUAVE database (some of the subjects were discarded due to random head movement and bad tracking) for the experiments and they were categorised into 8 groups as shown in Table I. Since the number of subjects are limited, a number of VVAD experiments have been performed according to the folds information as outlined in Table II to obtain an average result. For a particular fold, 75% of the data was selected for training of the GMM models and for the tuning of the segmentation thresholds and 25% was selected for testing.

C. Performance metrics calculation

In order to evaluate the performance of the VVAD system, performance metrics were designed to compare the final

TABLE I
SPEAKER LIST

Group	Speakers	Group	Speakers
I	s01, s02, s03	V	s19, s22, s23
II	s06, s07, s09	VI	s24, s25, s26
III	s10, s14, s15	VII	s27, s29, s30
IV	s16, s17, s18	VIII	s31, s32, s34

TABLE II
FOLDS FOR VVAD EXPERIMENTS

Fold	Training Groups	Testing Groups
1	I, II, III, IV, V, VI	VII, VIII
2	III, IV, V, VI, VII, VIII	I, II
3	I, II, V, VI, VII, VIII	III, IV
4	II,III, IV, V, VI, VII	I, VIII
5	I, II, III, IV, VII, VIII	V, VI
6	I, IV, V, VI, VII, VIII	II, III
7	II, III, V, VI, VII, VIII	I, IV
8	I, II, III, IV, VI, VIII	V, VII
9	I, II, III, V, VI, VII	IV, VIII
10	I, II, III, IV, V, VII	VI, VIII

speech segmentation files with the reference speech segmentation files as follows:

- Miss rate (MR) - How often a real speech frame is missed,

$$MR = \frac{T_m}{T_{ref}} * 100\%, \quad (1)$$

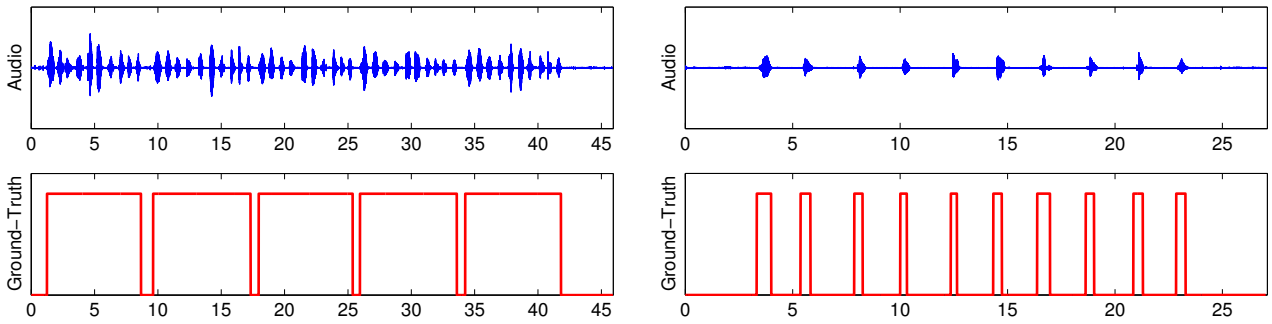
- False alarm rate (FAR) - How often a non-speech frame is detected as a speech frame,

$$FAR = \frac{T_{fa}}{T_{sys}} * 100\%, \quad (2)$$

- Half total error rate (HTER),

$$HTER = \frac{MR + FAR}{2} * 100\%, \quad (3)$$

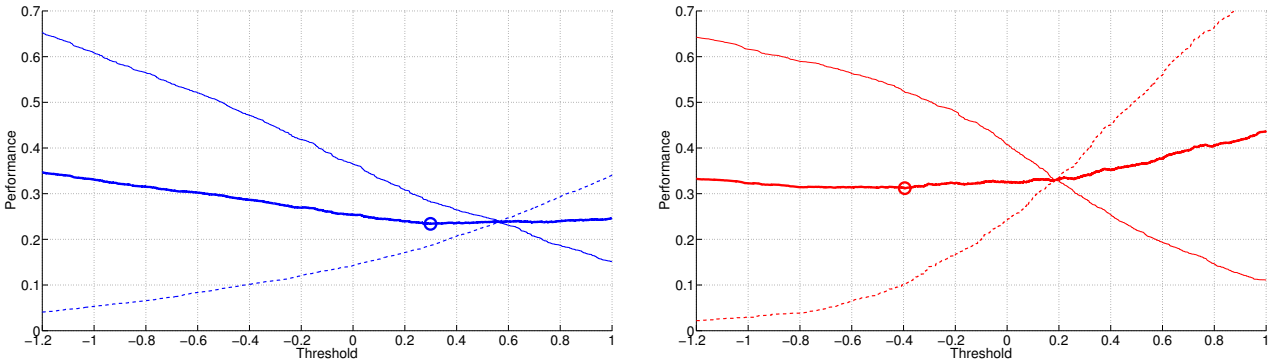
where, T_{fa} represents the duration of speech in false-alarm and T_{sys} represents the duration of speech in the system, T_m is defined as the duration of speech misses, and T_{ref} represents the reference event transcriptions.



(a) From frontal view.

(b) From profile view.

Fig. 6. Examples of the ground-truth values. The top rows illustrate the audio signals and the bottom rows illustrate how we derive the ground-truth values.



(a) From frontal view.

(b) From profile view.

Fig. 7. The variation of HTER performance with the threshold for the frontal and profile-view systems in a selected fold. The thick blue and red lines show the HTER for the frontal and profile views with the minimum HTER also indicated on the curves using a point. HTER define as the mean of the MR and FAR values. The solid thin lines are for the FAR and the dashed lines indicate the MR.

V. Experimental Results

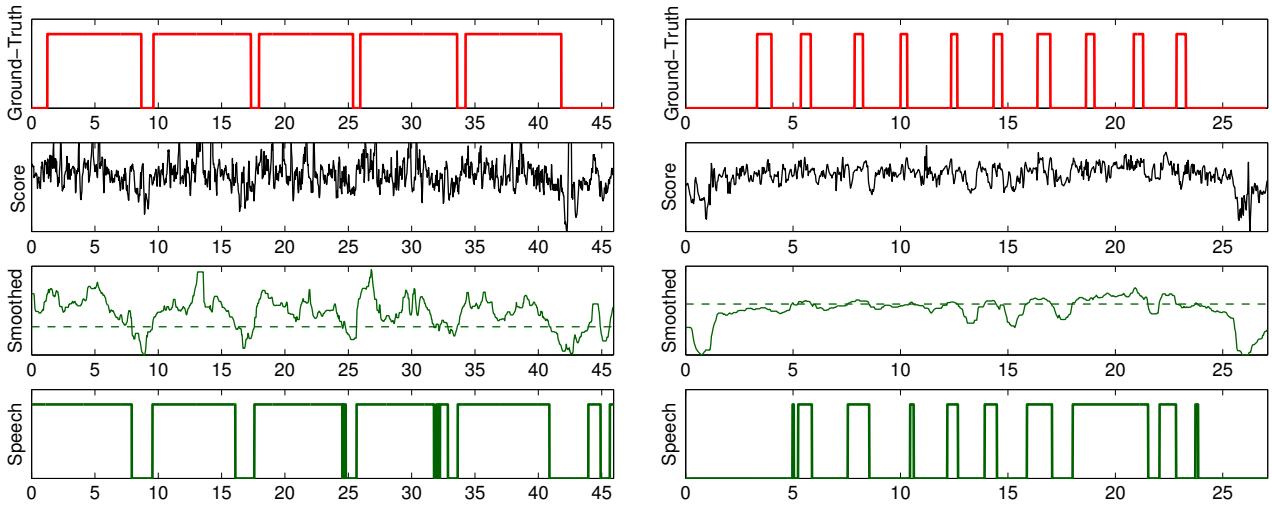
In this section, we report a number of experimental results on the performance of the developed VVAD system using the frontal, left and right views. The experiments were conducted using the CUAVE database which was described in the previous section. We report the experimental results using FAR, MR and HTER at segmentation thresholds based on the minimising the HTER over all speakers in the training fold. The choice of the segmentation threshold value is important in this framework as it separates the speech and the non-speech events. An example of threshold sensitivity for frontal and profile views systems is shown in Figure 7, which illustrates the variation of the HTER based on the threshold choice. The chosen operating point, based on minimising the HTER, on the curve is also indicated for the selected fold. Eight mixture GMMs were used and trained to minimise the HTER for all the experiments.

The performance of the frontal and the profile system are compared in Table III. As one would expect intuitively from the lesser amount of visual speech information that is present in the profile (side) views, the performance of the profile view based VVAD is less than that of the frontal view with HTER of 35.95% for the left profile and 33.95% for the right profile view compared of 25.9% for the frontal view. However, this results show that profile views are still capable of providing much of the visual modality to benefit VVAD. We view this result to be important in the development of efficient human computer interaction systems in many 'real-world' applications such voice based control in vehicular environment where frontal view of the speaker's (driver's) face may not always be available.

A typical example of the speech segmentation output obtained from the VVAD system using the frontal and the profile views is shown in Figure 8. In Figure 8 the third row presents the smoothed values from the framework with the threshold which is indicated as a dashed line. The smoothed value below the dashed line indicates the non-speech events and above indicates the speech events.

VI. Conclusion

In this paper, we have presented a voice activity detection framework using visual articulators. Specifically the paper addressed the effectiveness of the variabilities in the visual do-



(a) From frontal view.

(b) From profile view.

Fig. 8. Examples of VVAD on a sample of testing speakers which illustrates the score values in the second row according to the ground-truth in the first row, the smoothed values in the third row (dashed lines indicate the threshold value) and the output of the speech and non-speech events in the last row.

TABLE III
COMPARISON OF FRONTAL AND PROFILE VIEWS RESULTS

Performance metrics	Frontal results (%)	Left results (%)	Right results (%)
FAR	24.20	42.50	48.80
MR	27.60	29.40	19.10
HTER	**25.90**	**35.95**	**33.95**

main from the speaker's frontal and profile views for the task of VVAD. To our best knowledge this work represents the first attempt for VVAD using profile views. By our experiments, we demonstrated that profile views do contain important visual speech information, but as would be intuitively obvious, less compared with the frontal data due to the poor capturing of visual information from profile views. Having a VVAD system which can recognize the speech activity from both frontal and profile views will be a major benefit in the development of an efficient human computer interaction system based on audio-visual information.

ACKNOWLEDGMENT

The authors would like to thank Clemson University for freely supplying us the CUAVE database [11] for our research. This work was supported through the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

REFERENCES

[1] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a csp-based voice activity detector for distant-talking asr," *Proceedings of the EUROSPEECH 2003*, Geneva, 2003.

[2] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Transaction on Speech and Audio Processing*, vol. 11, pp. 498–505, 2003.

[3] L. Lamel, L. Rabiner, and A. Rosenberg, "An improved endpoint detector for isolated word recognition," *IEEE Trans Acoust, Voice Signal Processing*, pp. 777–785, 1981.

[4] J. Shen, J. Hung, and L. S. Lee, "Robust entropy based endpoint detection for voice recognition in noisy environment," *Proceedings of the 4th International Conference on Spoken Lanuage Processing*, pp. 881–884, 1996.

[5] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, V. V. Ollason, D. D. Povey, and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Entropic Ltd, 2002.

[6] P. Liu and Z. Wang, "Voice activity detection using visual information," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 609–612, 2004.

[7] T. Chen, Y. J. Hsu, X. Liu, and W. Zhang, "Principle component analysis and its variants for biometrics," *International Conference on Image Processing*, vol. 1, pp. 61–64, 2002.

[8] D. Sodoyer, B. Rivet, L. Girin, J. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2006.

[9] V. Libal, J. Connell, G. Potamianos, and E. Marcheret, "An embedded system for invehicle visual speech activity detection," in *Proceedings of the International Workshop on Multimedia and Signal Processing*, Chania, Greece, 2007, pp. 255–258.

[10] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," *IET Image Processing*, vol. 4, no. 6, pp. 463 –472, december 2010.

[11] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 2017–2020, Orlando, FL,USA ,2002.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition, 2001. CVPR 2001*, vol. 1, pp. 511–518, 2001.

[13] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," *IEEE 8th Workshop on Multimedia Signal Processing*, pp. 24 –28, oct. 2006.

[14] G. Potamianos, A. Verma, C. Neti, and S. Iyengar, G. Basu, "A cascade image transform for speaker independent automatic speechreading," *IEEE International Conference on Multimedia and Expo 2000, ICME 2000.*, vol. 2, pp. 1097–1100, 2000.

[15] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *in Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.